The Uganda Genome Project creates a roadmap for GWAS in Africa

Gurdasani D.*[1,2], Carstensen T.*[1,2] , Fatumo S.* [1,2,3], Chen G.[4], Franklin CS.*[1], Prado-Martinez J.* [1], Bouman H.* [1], Abascal F. [1,2] , Haber M. [1], Tachmazidou I. [1], Mathieson I.[5], Ekoru K.[1,6], DeGorter MK. [7], Nsubuga RN.[6], Finan C. [1], Wheeler E. [1], Chen L.[1], Cooper DN.[8], Schiffels S.[9], Chen Y. [1], Ritchie GRS. [1], Pollard MO. [1,2], Fortune MD. [1], Mentzer AJ.[10], Garrison E. [1], Bergström A.[1], Hatzikotoulas K.[1], Elding H. [1], Wain LV.[11], Ehret G.[12,13], Auer PL.[14], Kooperberg CL.[15] , Reiner AP.[16,17], Franceschini N.[18], Maher DP.[6], Trynka G. [1], Montgomery SB.[7,19], Kadie C.[20], Widmer C.[21], Xue Y.[1], Seeley J. [6,22], Asiki G. [6] , Kamali A. [6,19], Young EH. [1,2], Pomilla C. [1,2], Soranzo N. [1,23,24], Zeggini E. [1], Pirie F.[25], Morris AP.[26,10], Heckerman D.[20], Tyler-Smith C[1‡], Motala A.[25‡], Rotimi C[4‡], Kaleebu P. [‡5,21], Barroso I.[‡ 1], Sandhu MS.[1,2‡]

*joint authors
‡ equal contribution

Corresponding authors:

[1] Human Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
[2] Department of Medicine, University of Cambridge, Cambridge, UK
[3] H3Africa Bioinformatics Network (H3ABioNet) Node, National Biotechnology Development Agency, Federal Ministry of Science and Technology, Abuja, Nigeria
[4] Center for Research on Genomics and Global Health, National Institute of Health, USA
[5] Harvard Medical School, Department of Genetics, Boston MA, USA
[6] Medical Research Council/Uganda Virus Research Institute (MRC/UVRI) Uganda Research Unit on AIDS, Uganda
[7] Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA
[8] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK
[9] Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany
[10] The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[11] Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, UK
[12] McKusick-Nathans Institute of Genetic Medicine Johns Hopkins University School of Medicine Baltimore MD, USA
[13] Geneva University Hospitals, Rue Gabrielle-Perret-Gentil,41211 Genève 14, Switzerland
[14] Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee WI, USA
[15] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle WA, USA
[16] Department of Epidemiology, University of Washington, Seattle WA, USA
[17] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle WA, USA
[18] Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA
[19] Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
[20] Microsoft Research, Redmond, CA, USA
[21] Microsoft Research, Los Angeles, CA, USA
[22] London School of Hygiene and Tropical Medicine, London, UK
[23] Department of Haematology, University of Cambridge, Cambridge, UK
[24] The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics, University of Cambridge, Cambridge, UK
[25] Department of Diabetes and Endocrinology, University of KwaZulu-Natal, Durban, South Africa
[26] Department of Biostatistics, University of Liverpool, Liverpool, UK

Manjinder Sandhu ms23@sanger.ac.uk
Inês Barroso ib1@sanger.ac.uk
Charles Rotimi rotimic@mail.nih.gov
Ayesha Motala MOTALA@ukzn.ac.za
Chris Tyler-Smith cts@sanger.ac.uk
Pontiano Kaleebu pontiano.kaleebu@mrcuganda.org

Abstract

Genomic studies in African populations provide unique opportunities to understand disease aetiology, human genetic diversity and population history in a regional and a global context. In the largest study of its kind to date, comprising genome-wide data from 6,400 individuals from rural Uganda, and including whole-genome sequence from 1,978 individuals, we find evidence of geographically correlated fine-scale substructure, as well as complex admixture from eastern African hunter-gatherer and Eurasian populations. Examining 34 cardiometabolic traits, we demonstrate systematic differences in trait heritability between European and African populations, probably reflecting the differential impact of genetic and environmental factors on traits. In the first multi-trait pan-African GWAS, we identify 12 novel loci associated with anthropometric, liver, haematological, lipid and glycemic traits. Our findings suggest that several functionally important signals at known and novel loci may be driven by differentiated variants within and specific to Africa, influencing reproducibility of associations signals across populations; these findings have important implications for the design of medical genetics research in Africa and globally. We highlight the value of the largest sequence panel from Africa to date as a global resource for population genetics, imputation and understanding the mutational spectrum and its clinical relevance in African populations. Alongside phenotype data, we provide a rich new genomic resource for researchers in Africa and globally.

Introduction

Africa is central to our understanding of human origins, genetic diversity and disease susceptibility.[1] The marked genomic diversity and allelic differentiation among populations in Africa, in combination with the substantially lower linkage disequilibrium (correlation) among genetic variants, has the potential to provide new opportunities to understand disease aetiology relevant to African populations but also globally.[1,2] Consequently, there is a clear scientific and public health need to develop large-scale efforts that examine disease susceptibility across diverse populations within the African continent. Such efforts will need to be fully integrated with research-capacity-building initiatives across the region.[3]

Countries in Africa are undergoing epidemiological transitions—with a high burden of endemic infectious disease and growing prevalence of non-communicable diseases.[4] Importantly, because of varying environments, population history and demography, and adaptive evolution, the spectrum and distribution of risk factors for a broad range of cardiometabolic and infectious diseases, and their individual contribution, may differ among populations globally.[5] The study of population specific genetic variation and differences in allelic frequencies among populations, due to either selection or genetic drift, provides a distinct approach to identify novel disease susceptibility loci. However, despite the value of conducting such studies in Africa, there have been relatively few investigations of population diversity or the genetic determinants of cardiometabolic or infectious traits and diseases across the region.

To conduct genetic studies in diverse populations across Africa, appropriate study designs that take into account population structure, admixture and genetic relatedness (overt and cryptic), as well as the

development of genetic tools to capture variation in African genomes, are needed.[2] To leverage the relative benefits of different strategies, we undertook a combined approach of genotyping and low coverage whole-genome sequencing (WGS) in a population-based study of 6,400 individuals from a geographically defined rural community in South-West Uganda (**Figure 1a, Supplementary Information 1.0, Supplementary Figures 1-5** and **Supplementary Tables 1-2**). We present data from 4,778 individuals with genotypes for ~2.2 million SNPs from the Uganda genome-wide association study (UGWAS) resource (**Supplementary Information 1.1-1.3**), and sequence data (**Supplementary Information 1.4, Supplementary Figures 2-4**) on up to 1,978 individuals spanning 41.5M SNPs and 4.5M indels (UG2G) ; 343 individuals overlap between the two datasets (**Figure 2** and **Supplementary Information 1.4**). Collectively, these data represent the Uganda Genome Resource (UGR). To enhance discovery, we also include collective data on up to 14,126 individuals from across the African continent for genome wide association analysis (**Supplementary Information 4.0**).

Using these resources, we conducted a series of analyses to: 1) understand the population structure and demographic history in a geographically-defined population from Uganda (**Supplementary Information 2.0**); 2) refine estimates of heritability of 34 complex traits, accounting for environmental correlation among individuals (**Supplementary Information 3.0**); 3) assess the spectrum of genetic variants associated with cardiometabolic and other complex traits in populations from sub-Saharan Africa (**Supplementary Information 4.0**)**;** 4) describe the spectrum of disease-causing mutations in the UG2G cohort (**Supplementary Information 5.0**); and) highlight the value of the UG2G sequence panel as an imputation resource (**Supplementary Information 5.0)**. Importantly, the UGR was designed to help develop local resources for public health and genomic research, including building research capacity, training and collaboration across the region. We envisage that data from these studies will provide a global resource for researchers, as well as facilitate genetic studies in African populations.

Results

Population structure and demographic history in a rural Ugandan community

To help inform strategies that account for genetic diversity, and develop resources to reliably capture the broad allelic spectrum in a geographically-defined rural Ugandan community, we provide the first detailed description of genetic diversity and fine-scale structure among nine ethno-linguistic population groups from the UGR (**Supplementary Information 2.0**).

Although the study population represents a geographically-defined rural community from the Kalungu District in South-West Uganda (**Figure 1a**), principal components (PC) inferred from fineSTRUCTURE showed evidence suggestive of population substructure (**Figure 1b**, **Supplementary Information 2.0** and **Supplementary Figures 5-14, Supplementary Table 3**), with clines along PC1 and PC2 being highly correlated with Eurasian and East African Nilo-Saharan ancestry, respectively (**Supplementary Figures 10-12** and **Supplementary Information 2.2.2**). Using fineSTRUCTURE[6] and Procrustes analyses, we show for the first time that substructure among ethno-linguisitic groups in a rural Ugandan community is correlated with their historical geographical origins (**Supplementary Tables 4-6, Supplementary Figures 13-17**). Populations from the central region of Uganda (the Baganda, Basoga and Batooro), migrant populations from Rwanda, Burundi, Tanzania and those from South-western Uganda (Bakiga, Banyankole and Bafumbira) form separate clades (**Supplementary Information 2.2-2.3,**

Supplementary Figure 5 and **Supplementary Figures 13-14**). This is consistent with recent patterns of population migration into Uganda from neighbouring regions and border countries, including from Rwanda and Burundi (**Supplementary Information 2.1**).[7]

PCA, unsupervised ADMIXTURE, and fineSTRUCTURE analysis were suggestive of broad Eurasian and hunter-gatherer admixture across ethno-linguistic groups (**Figures 1c-d**, **Supplementary Information 2.0**, **Supplementary Figures 18-24)** We use multiple lines of evidence, including formal tests for admixture (f_3, f_4 tests, MALDER),[8 6] the double conditioned site frequency spectrum, and the distribution of mitochondrial and Y chromosome haplogroups to identify Eurasian gene flow in Uganda (**Supplementary Tables 9-22**, **Supplementary Figures 25-30** and **Supplementary Information 2.4.2-2.4.7**). Using the Conditional Random Field model (CRF), we show strong evidence of Neanderthal ancestry in Uganda, providing strong evidence of Eurasian admixture resulting from back to Africa migration. (**Supplementary Table 15** and **Supplementary Information 2.4.5**). Our findings suggest that the patterns and extent of this admixture varied among these regional populations. We show evidence of extensive Eurasian admixture across all populations with multiple events ranging between 200 and 9,000 years ago (ya), consistent with back migration into Africa[9 10](**Figure 1d** and **Supplementary Information 2.4.2-2.4.7**). We also observed novel widespread hunter-gatherer admixture in Uganda, including strong admixture signals from co-regional hunter-gatherer populations such as the rainforest hunter-gatherers and Hadza (**Figure 1c**, **Supplementary Information 2.4.3, 2.4.7**, **Supplementary Tables 10-11 and Supplementary Table 22**). These signals were specific to the Ugandan and other East African populations (**Figure 1e** and **Supplementary Table 11, Supplementary Table 22**), and spanned between 2,400 and 4,500 ya (**Figure 1d**)—suggesting assimilation of hunter-gatherer ancestry through population migration events, including the Bantu expansion. We also provide direct evidence of the temporal shape of this ancient admixture, reflected by a relative excess of allele sharing with ancient Neolithic Europeans (both early European farmers[11] and their Anatolian ancestors[12]), consistent with previous reports,[9,13] and novel evidence for regionally specific shared ancestry with an ancient East African genome (Mota[5]) in the Ugandan populations (**Supplementary Information 2.4.7** and **Supplementary Tables 18-22**).

Using MSMC2 on high coverage genomes (**Supplementary Information 2.5**),[14] we show that the Ugandan populations follow a similar demographic history to other Bantu speaking populations, with the Ugandan population split from Yoruba, Nigeria (YRI) ~11,500 ya, with ongoing gene flow between Uganda-LWK in recent times (**Supplementary Information 2.5, Supplementary Tables 23-24, Supplementary Figures 31-35**). The Uganda-YRI divergence is older than predicted by the Bantu expansion. [15] However, these differences could also reflect population differences, including varying patterns of Eurasian and regional admixture in East and West African populations. It also should be noted that these divergence times are lower bounds, and are likely to be affected by gene flow between these populations following divergence, as has been previously documented. [14]

We then explored recent population history by examining rare variant sharing between the Baganda and other populations; we examined variants occurring only twice in the entire dataset (designated f_2) (**Figure 1e and Supplementary Information 2.3**). Dating haplotypes surrounding f_2 variants can provide important information about the interrelation among populations, including ancient and recent population divergence.[16] Using this approach, we find that f_2 variants shared between European and

Ugandan populations are more recent than those shared between European and West African populations (median f_2 dates were ~19,500 ya for Baganda compared with ~51,000 ya for YRI). This finding is consistent with back migration[10] and Eurasian admixture in the Uganda populations (**Supplementary Information 2.3, Supplementary Table 8** and **Supplementary Figures 19-20**).[2,13] Examining Ugandan populations in the context of other African populations, we find that f_2 sharing between Ugandan populations and Ethiopians tend to be older (median f_2 dating was ~23,000 ya) than Ugandan-West African splits, probably reflecting a combination of deeper population splits between Bantu- and Afro-Asiatic-speaking groups, and relatively high Eurasian admixture in the Ethiopian populations. We also find evidence of very ancient divergence (with a median f_2 dating of ~29,000 ya) between Baganda and Zulu (**Figure 1e** and **Supplementary Figure 20**); this could reflect old f_2 sharing with Khoe-San haplotypes present among Zulu and other Southern African populations. Our large African sequence resource allows the first such examination of shared rare variation among populations, and highlights the complex demographic histories of populations in this region.

Heritability of cardiometabolic traits in a rural Ugandan community

Narrow-sense heritability represents the fraction of phenotypic variation in a population that is due to additive genetic variation. As such, it represents an important metric determining the genetic basis of complex traits and diseases. However, heritability of complex traits in African populations is not known. We assessed heritability for 34 complex cardiometabolic traits using a newly developed mixed model approach that also models environmental correlation[17] (**Figure 3** and **Supplementary Information 3.0**).

We found marked variation in heritability across traits in UGWAS (**Figure 3**, **Supplementary Information 3.0** and **Supplementary Table 25**), and find clear statistical differences in heritability estimates for several traits, compared to European populations (**Figure 3** and **Supplementary Tables 26-28**). For example, the narrow-sense heritability for height was 49% in UGWAS, contrasting with estimates of 70-80% in European populations ($p<0.0001$) (**Figure 3, Supplementary Tables 26-28**). We speculate that these differences may be due to varying patterns of genetic loci influencing height or other traits in European and African populations, or perhaps more plausibly due to a larger proportion of environmental variation explaining phenotypic variance (e.g. under-nutrition in rural African populations may attenuate the effects of genetic variation on height).[18] We also found evidence for statistical gene-environment interaction for waist-hip ratio, red blood cell distribution width and haematocrit (permutation $p=<0.0001$). These statistical interactions may represent interplay between genetic factors and dietary factors, iron stores and nutritional status (**Supplementary Table 25**). Reliable assessment of the interrelation between genetic and environmental variation, including specific environmental indices, will require application of these methods to much larger-scale studies with relevant phenotypic information. Examining locus-specific heritability would complement direct assessments of population differences in heritability of population traits.

GWAS of cardiometabolic traits in African populations

To assess the spectrum of genetic variants associated with cardiometabolic traits in African populations, we performed a GWAS of 34 cardiometabolic traits in up to 14,126 individuals from across the African continent, including populations from Ghana, Kenya, Nigeria, South Africa and Uganda (**Supplementary Information 4.0**, **Supplementary Data Table 1, Figure 4** and **Table 1**). We also

undertook an exploration of analytical strategies for populations with marked genetic diversity and structure to inform GWAS design and analytical strategies in African populations (**Supplementary Information 4.6-4.10**, **Supplementary Tables 31**, **Supplementary Figures 38-39)**. To maximise opportunities for genomic discovery, we meta-analysed GWAS data from all study populations using the Han-Eskin random-effect meta-analytic approach implemented in METASOFT[19] (**Supplementary Information 4.8**). We first re-assessed thresholds for genome-wide statistical significance in African populations using several approaches[20-23] and found that a statistical threshold of 5.0×10^{-9} is more relevant in populations with high genetic diversity and relatively lower levels of LD (**Supplementary Information 4.9-4.10**). In our discovery meta-analysis, we identified 41 distinct association signals at this new genome-wide statistical threshold for at least one trait (**Supplementary Information 4.13**, **Supplementary Data Table 1, Supplementary Figures 40-66, Supplementary Tables 32-39**). More than half of these distinct association signals (23/41) were attributable to genetic variants specific to African populations (monomorphic or extremely rare in Europeans) (**Supplementary Information 4.13 and Supplementary Data Table 1**), including new distinct signals at previously identified loci (**Table 1**). Twelve association signals were novel loci—trait assocations, including for liver function traits, anthropometric indices, HbA1c, and heamatological and blood cell traits (**Table 1, Figure 4** and **Supplementary Data Table 1**) . We note that most novel discovery was driven by low frequency (0-5% MAF) variants; ten of these 12 novel association signals were low frequency or monomorphic (MAF<2%) in European populations (**Supplementary Data Table 1**). Our findings demonstrate the potential for novel discovery in African GWAS. Collectively, these findings provide the first empirical evidence to support theoretical models that suggest that power for discovery increases in meta-analyses of ethnically diverse populations, specifically driven by increased detection of low frequency and population-specific novel associations.[24]

Our novel association signals included a functionally relevant association between a 3.8Kb deletion (-α3.7), known to cause alpha thalassemia, and HbA1C levels at $p = 2.5^{-17}$ (**Supplementary Information 4.13, Table 1** and **Figure 4**). The -α3.7 variant is thought to have risen to high frequencies in African populations in regions endemic for malaria by virtue of providing resistance to severe malaria.[25] Our findings recapitulate the need to more fully understand functional variation, including for heamoglobinopathies, that may explain a substantial proportion of variation in HbA1c in African populations, as they may have a direct impact on the utility of using HbA1C as a clinical tool for detection and diagnosis of diabetes in Africa.[26] We also discovered an association signal at the glutamate-pyruvate transaminase (*GPT*) locus with the liver function biomarker amino-transferase (ALT), $p = 5.7^{-38}$ (**Supplementary Information 4.13, Table 1** and **Figure 4**). The lead SNP was African population-specific variant lying in a *GPT* splice site region. *GPT* is the gene primary responsible for the production of the ALT enzyme, highlighting the biological relevance of this association signal.

Notably, we found limited reproducibility among several functionally important known and novel loci in our meta-analysis (e.g. the *DARC* locus associated with monocyte count). Some of these associations appear to be driven by population specific rare or low frequency variants (**Supplementary Information 4.13 and Supplementary Data Table 2**). We also found evidence of statistical heterogeneity of effect in associations with several traits (**Supplementary Information 4.13, Supplementary Data Table 1 and Supplementary Data Table 2**); this was found to arise from differences in LD structure around causal or lead variants or presence of multiple distinct variants at the locus (alleleic heterogeneity), providing

important insights into the genetic architecture of traits (**Supplementary Information 4.13**). Given high genetic diversity, and regionally specific patterns of admixture, we highlight the need to design GWAS studies to account for these differences in allele frequency spectrum, and LD patterns across the African cohorts. With caveats for rare variant discovery in some scenarios (**Supplementary Information 4.13**), our analyses emphasize the value of utilizing diverse populations across the region—to maximise opportunities for genomic discovery in the region.

A whole genome resource for population and medical genetics

With the largest whole genome sequence dataset from Africa to date (**Figure 2** and **Supplementary Information 5.0**), with nearly a quarter of all variants being novel relative to currently available genome sequence data, we present a unique resource to examine the spectrum of human genetic diversity in Africa, regional and global population history, and the phenotypic consequences of putative functional variants among individuals, as well as a resource to facilitate medical genetics studies in the region.

As expected, and consistent with the out-of-Africa model, Africans carry the largest number of variants, the overwhelming majority being rare (**Supplementary Information 5.0, Supplementary Tables 40-41, Supplementary Figures 67-70**). In line with these observations, African populations provide greater opportunities for variant discovery as a function of sample size (**Supplementary Information 5.3 and Supplementary Figure 71**). We identified 9.5 M novel variants in the UG2G resource that are not present in the 1000G Phase 3, AGVP and UK10K data (**Figure 2**), highlighting the importance of assessing diverse populations on a larger scale.

We also explored the putative functional consequences of variation in the UG2G population (**Supplementary Information 5.4, Supplementary Figures 72-75, Supplementary Tables 42-43**). Consistent with overall diversity, UG2G participants carried more missense variants per individual compared with the UK10K population (12,198 and 10,153 variants/individual respectively). However, missense variants formed a greater proportion of all variation among Europeans. (**Supplementary Figure 72**). For disease-causing mutations (DMs), as annotated by the HGMD (**Supplementary Information 5.4** and **Supplementary Figures 73-74**), we identified a median of 29 DMs/individual in our cohort compared to 25 DMs/individual in UK10K, despite more extensive studies in European populations, and potentially biased ascertainment.[27] By contrast, in UG2G, we observed a median of 3 homozygous DMs/individual compared with to 4 homozygous DMs/individual in UK10K (**Supplementary Information 5.4**) ($p<2X10^{-16}$). The distribution of the mutational spectrum in African and European populations is consistent with previous reports, [28,29] and the impact of differences in demographic history among these populations (an evolutionary bottleneck in European populations with subsequent drift).[28]

Allelic frequency differences between populations along with clinical phenotype data may provide insights into the functional relevance of putative DMs. On assessing 47 DMs that were common in our cohort (MAF>5%) but rare or absent in the UK10K data (**Supplementary Table 43**),[30] we identified established causal loci associated with haematological traits, such as the *G6PD* and sickle cell variants (*HBB*), which are common in UG2G, but absent from the UK10K data, consistent with these loci being

under positive or balancing selection and protective against malaria (**Supplementary Table 43**). However, we also show that several putative DMs that are common in UG2G but rare in UK10K show no strong evidence for association with relevant cardiometabolic traits, indicating that they are unlikely to be disease-causing or have different or lower penetrance within African populations due to complex factors, including epistasis, or gene-environment interplay (**Supplementary Figure 75**). This emphasises the need to carefully evaluate the impact of putative functional or disease-causing mutations because they may not have any clinical or biological relevance, or be readily transferable across populations.[27,31]

Finally, we assess the impact of the addition of the UG2G panel to existing reference panels on imputation accuracy among populations from Sub-Saharan Africa (**Supplementary Information 5.5**). We show that addition of the UG2G panel to existing sequence panels with African haplotypes, such as the 1000 Genomes Phase 3 panel, and the African genome variation sequence panel (combined n=3,895), markedly improved imputation accuracy (r^2 increase by 0.08 (MAF<=0.01) and 0.04 (all MAF)) for rare and common variants in Ugandan populations (**Figure 5, Supplementary Figure 76** and **Supplementary Information 5.4**). Additionally, we observe a substantial increase in imputation accuracy across the allele frequency spectrum generally in East African populations, including Nilo-Saharan linguistic groups such as the Kalenjin (**Figure 5**), probably reflecting haplotype sharing across the region. The number of variants successfully imputed (info≥0.3) substantially increased using the UG2G panel in comparison with the 1000 Genomes phase III and AGVP combined panels, with an additional 8M variants being accurately imputed in Baganda, and 1.5M additional variants imputed accurately among other East African populations (**Figure 5**). This resource coupled with current initiatives to generate very large reference panels (e.g. the Haplotype Consortium[32]) has important implications for facilitating genetic studies in this region.

Discussion

We demonstrate the value of cataloging whole-genome sequence variation and undertaking GWAS of cardiometabolic traits in a geographically-defined rural community in south-west Uganda. The Uganda Genome Project datasets provide rich genomic resources for studies of human population history and genome-wide association studies, and provide a mechanism to evaluate the clinical relevance of genetic diversity both in African populations and globally.

We present evidence for fine-scale structure and admixture in this Ugandan community, reflecting complex ancient and recent population migrations and expansions in East Africa. Our findings highlight the need for larger-scale deep sequencing, including a systematic assessment of hunter-gatherer populations across Africa, to more fully understand the genetic history and diversity of Africa. Sequencing of DNA from ancient skeletal material across Africa will greatly facilitate such efforts[33]—allowing stronger inferences into the source of genetic diversity and population history in Africa and globally.

Accounting for environmental correlation, we describe statistical differences in heritability for traits between African and European populations; these may be suggestive of differences in the interplay between genetic and environmental effects on heritable traits, as well as the impact of differences in genetic architecture as a result of selection, drift and historical demographic events. Our findings re-

iterate the dynamic and context-specific nature of heritability, potentially varying among populations, demographic factors and environmental exposures.[34]

Our identification of several novel susceptibility loci across a range of complex traits argues for scaling efforts in the region. The continental and population-specificity of a large proportion of these association signals suggests that inclusion of diverse populations across Africa in GWAS may have the greatest potential for discovery of novel targets. Furthermore, understanding differences in heritability, and identifying the full spectrum of genetic variation associated with complex traits and diseases across Africa, will require much larger-scale prospective studies that should include rich genomic and phenotypic data for complex traits and diseases, as well as information on environmental factors. In these contexts, our results provide a framework for undertaking more extensive GWAS in populations from Africa.

Since genetic diversity is greatest in African populations, including a substantial proportion of genetic variation that is continentally and regionally distinct, it will be critical to understand the functional and biological relevance of this diversity. This has global implications. Understanding the biological basis for population-specific association signals, as well as the impact and tranferability of putatively functional and disease causing mutations at the individual and population level, will require representative genomic resources—emphasizing the need for the parallel development of transcriptomic and cellular biological resources at the population level that reflect global human diversity.[35]

Acknowledgements
This work was funded by the Wellcome Trust, The Wellcome Trust Sanger Institute (WT098051), the UK Medical Research Council (G0901213-92157, G0801566, and MR/K013491/1), and the Medical Research Council/Uganda Virus Research Institute Uganda Research Unit on AIDS core funding. This work was funded in part by IAVI with the generous support of the United States Agency for International Development (USAID) and other donors. The full list of IAVI donors is available at http://www.iavi.org. The contents of this manuscript are the responsibility of IAVI and co-authors and do not necessarily reflect the views of USAID or the US Government. We thank the African Partnership for Chronic Disease Research (APCDR) for providing a network to support this study as well as a repository for deposition of curated data. We thank all study participants who contributed to this study. We also acknowledge the National Institute for Health Research Cambridge Biomedical Research Centre. The authors wish to acknowledge the use of The Uganda Medical Informatics Centre (UMIC) compute cluster. Computational support from UMIC was made possible through funding from the Medical Research Council (MC_EX_MR/L016273/1). We acknowledge the Sanger core pipeline teams for their help with sequencing and mapping the whole genome sequence data. APM is a Wellcome Senior Fellow of Basic Biomedical Science (under award WT098017). NS's research is supported by the Wellcome Trust (Grant Codes WT098051 and WT091310), the EU FP7 (EPIGENESYS Grant Code 257082 and BLUEPRINT Grant Code HEALTH-F5-2011-282510) and the National Institute for Health Research Blood and Transplant Research Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge in partnership with NHS Blood and Transplant (NHSBT). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or NHSBT. AJM was funded by the Wellcome Trust (WT106289). We acknowledge use of summary data from the Global Lipids Genomics Consortium (GLGC). [36]

We acknowledge the H3Africa Bioinformatics Network (H3ABioNet) Node, National Biotechnology Development Agency (NABDA), Federal Ministry of Science and Technology (FMST) Abuja, Nigeria for funding SF for his post-doctoral research. DNC wishes to acknowledge the financial support of Qiagen Inc through a License Agreement with Cardiff University. We also acknowledge the 1000 Genomes Project, UK10K, Simon's Foundation Genome Diversity Project and African Genome Variation Project (AGVP) for providing data resources that were used to contextualise the UG2G data. The GATK3 program was made available through the generosity of Medical and Population Genetics program at the Broad Institute, Inc.

Data sharing

Summary GWAS and allele frequency data are publicly available at http://www.apcdr.org/data/. All individual level data, phenotype, genotype and sequence data are available under managed access to researchers. Requests for access to the phenotypic data will be granted for all research consistent with the consent provided by participants. This would include any research in the context of health and disease, that does not involve identifying the participants in any way. The APCDR committees are responsible for curation, storage, and the independent APCDR committee is responsible for the sharing of phenotypic and genetic data under managed access. The array and sequence data have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/, study/dataset accession numbers EGAS00001001558/EGAD00010000965 and EGAS00001000545/EGAD00001001639, respectively) and can be requested through datasharing@sanger.ac.uk. Requests for access to phenotype data may be directed to data@apcdr.org. While data cannot be released on public databases as this would conflict with the study protocol and participant consent under which data were collected, we aim to facilitate data access for all bona fide researchers. Applications are reviewed by an independent data access committee (DAC) and access is granted if the request is consistent with the consent provided by participants. The data producers may be consulted by the DAC to evaluate potential ethical conflicts. Requestors also sign an agreement which governs the terms on which access to data is granted.

Bibliography

1 Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-1044, doi:10.1126/science.1172257 (2009).

2 Gurdasani D., C. T., Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard M O, Choudhury A, Ritchie G R S, Xue Y, Asimit J, Nsubuga R N, Young E H, Pomilla C, Kivinen K, Rockett K , Kamali A, Doumatey A P, Asiki G, Seeley J, Sisay-Joof F , Jallow M, Tollman S, Mekonnen E, Ekong R, Oljira T, Bradman N, Bojang K, Ramsay M, Adeyemo A, Bekele E, Motala A, Norris S A, Pirie F, Kaleebu P, Kwiatkowski D, Tyler-Smith C, Rotimi C, Zeggini E and Sandhu M S. The African Genome Variation Project shapes medical genetics in Africa. *Nature* (2014).

3 Consortium, H. A. Research capacity. Enabling the genomic revolution in Africa. *Science* **344**, 1346-1348, doi:10.1126/science.1251546 (2014).

4 Organisation, W. H. Health Transition. (2015).

5 Campbell, M. C. & Tishkoff, S. A. The evolution of human genetic and phenotypic variation in Africa. *Curr Biol* **20**, R166-173, doi:10.1016/j.cub.2009.11.050 (2010).
6 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453, doi:10.1371/journal.pgen.1002453 (2012).
7 Richards, A. I. *Economic development and tribal change: a study of immigrant labour in Buganda*. (W. Heffer and Sons, 1954).
8 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093, doi:10.1534/genetics.112.145037 (2012).
9 Gallego Llorente, M. *et al.* Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* **350**, 820-822, doi:10.1126/science.aad2879 (2015).
10 Henn, B. M. *et al.* Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* **8**, e1002397, doi:10.1371/journal.pgen.1002397 (2012).
11 Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413, doi:10.1038/nature13673 (2014).
12 Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, doi:10.1038/nature16152 (2015).
13 Pickrell, J. K. *et al.* Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci U S A* **111**, 2632-2637, doi:10.1073/pnas.1313787111 (2014).
14 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).
15 de Filippo, C., Bostoen, K., Stoneking, M. & Pakendorf, B. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proc Biol Sci* **279**, 3256-3263, doi:10.1098/rspb.2012.0318 (2012).
16 Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet* **10**, e1004528, doi:10.1371/journal.pgen.1004528 (2014).
17 Heckerman, D. *et al.* Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc Natl Acad Sci U S A* **113**, 7377-7382, doi:10.1073/pnas.1510497113 (2016).
18 Nalwoga, A. *et al.* Nutritional status of children living in a community with high HIV prevalence in rural Uganda: a cross-sectional population-based survey. *Trop Med Int Health* **15**, 414-422, doi:10.1111/j.1365-3156.2010.02476.x (2010).
19 Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *American Journal of Human Genetics* **88**, 586-598, doi:10.1016/j.ajhg.2011.04.014 (2011).
20 Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* **32**, 361-369, doi:10.1002/gepi.20310 (2008).
21 Chen, Z. & Liu, Q. A new approach to account for the correlations among single nucleotide polymorphisms in genome: wide association studies. *Hum Hered* **72**, 1-9, doi:10.1159/000330135 (2011).
22 Moskvina, V. & Schmidt, K. M. On multiple-testing correction in genome-wide association studies. *Genet Epidemiol* **32**, 567-573, doi:10.1002/gepi.20331 (2008).
23 Nyholt, D. R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* **74**, 765-769, doi:10.1086/383251 (2004).

24 Pulit, S. L., Voight, B. F. & de Bakker, P. I. Multiethnic genetic association studies improve power for locus discovery. *PLoS One* **5**, e12600, doi:10.1371/journal.pone.0012600 (2010).
25 Mockenhaupt, F. P. *et al.* Alpha(+)-thalassemia protects African children from severe malaria. *Blood* **104**, 2003-2006, doi:10.1182/blood-2003-11-4090 (2004).
26 Herman, W. H. & Cohen, R. M. Racial and ethnic differences in the relationship between HbA1c and blood glucose: implications for the diagnosis of diabetes. *J Clin Endocrinol Metab* **97**, 1067-1072, doi:10.1210/jc.2011-1894 (2012).
27 Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet* **91**, 1022-1032, doi:10.1016/j.ajhg.2012.10.015 (2012).
28 Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* **47**, 126-131, doi:10.1038/ng.3186 (2015).
29 Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-997, doi:10.1038/nature06611 (2008).
30 Consortium, U. K. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90, doi:10.1038/nature14962 (2015).
31 Saraf, S. L. *et al.* Differences in the clinical and genotypic presentation of sickle cell disease around the world. *Paediatr Respir Rev* **15**, 4-12, doi:10.1016/j.prrv.2013.11.003 (2014).
32 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*, doi:10.1038/ng.3643 (2016).
33 Pickrell, J. K. & Reich, D. Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet* **30**, 377-389, doi:10.1016/j.tig.2014.07.007 (2014).
34 Haworth, C. M. & Davis, O. S. From observational to dynamic genetics. *Front Genet* **5**, 6, doi:10.3389/fgene.2014.00006 (2014).
35 Chang, E. A. *et al.* Derivation of Ethnically Diverse Human Induced Pluripotent Stem Cell Lines. *Sci Rep* **5**, 15234, doi:10.1038/srep15234 (2015).
36 Consortium, G. L. G. Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-1283, doi:10.1038/ng.2797 (2013).